Translational systems pharmacology-based predictive assessment of drug-induced cardiomyopathy

Dimitris E Messinis[1,2], Ioannis N Melas[1,3], Junguk Hur[4], Navya Varshney[5], Leonidas G Alexopoulos[6], and Jane PF Bai[1,§]

1. Office of Clinical Pharmacology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland 20993, USA
2. Present address: School of Mechanical Engineering, National Technical University of Athens, Zografou 15780, Greece
3. Present address: Translational Bioinformatics, UCB Celltech, 208 Bath Road, Slough, SL1 3WE UK
4. Department of Biomedical Sciences, University of North Dakota, School of Medicine & Health Sciences, Grand Forks, North Dakota 58502, USA
5. College of Pharmacy, University of Michigan, 428 Church St, Ann Arbor, Michigan, 48109, USA
6. School of Mechanical Engineering, National Technical University of Athens, Zografou 15780, Greece
§. Corresponding Author

Corresponding author[§]: Office of Clinical Pharmacology, Office of Translational Science, CDER, FDA, Silver Spring, Maryland. Tel: 301-796-2473; jane.bai@fda.hhs.gov

Conflict of Interest: The authors declared no competing interests for this work.

Funding: Dimitris Messinis was supported as ORISE fellow through a 2013 Medical Counter Measures Grant to JPFB. Navya Varshney was supported as a pharmacy intern via fellowship from Office of Clinical Pharmacology. Assistance on using Betsy and computation efficiency by Dr. Mike Mikailov (computer scientist, CDRH) is appreciated. This project was supported by an appointment to the Research Participation Program at CDER, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy and the FDA. Helpful discussion about elastic net with Dr. Hui Zou (Professor of Statistics, University of Minnesota) is appreciated. Comments on cardiac contractility and cardiomyopathy by Dr. David Strauss (Division Director, Division of Applied Regulatory Science, Office of Clinical Pharmacology) are appreciated. This project was supported by an appointment to the Research Participation Program at CDER, administered by

the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy and the FDA.

Disclaimer: This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Abstract

Drug-induced cardiomyopathy contributes to drug attrition. We compared two pipelines of predictive modeling: 1) applying elastic net (EN) to differentially expressed genes (DEGs) of drugs; 2) applying integer linear programming (ILP) to construct each drug's signaling pathway, starting from its targets to downstream proteins, to transcription factors, and to its DEGs in human cardiomyocytes, and then subjecting the genes/proteins in drugs' signaling networks to elastic net regression. We classified 31 drugs with availability of DEGs into 13 toxic and 18 non-toxic drugs based on a clinical cardiomyopathy incidence cut-off of 0.1%. ILP-augmented modeling increased prediction accuracy from 79% to 88% (sensitivity: 88%; specificity: 89%) under leave-one-out cross validation. ILP-constructed signaling networks of drugs were better predictors than DEGs. Per literature, the microRNAs that reportedly regulate expression of our 6 top predictors are of diagnostic value for natural heart failure or doxorubicin-induced cardiomyopathy. This translational predictive modeling might uncover potential biomarkers.

Key words: drug-induced cardiomyopathy, differentially expressed genes, drug targets, protein interaction, microRNAs, and predictive modeling.

Introduction

Serious and life threatening drug-induced adverse events cause drug attrition at various stages of drug development or modification of treatment regimens. For instance, anthracyclines, though effective to treat cancers, are known to cause irreversible, dose-dependent cardiotoxicity (contractility-related toxicity)[1]. Most recently, targeted therapy with tyrosine kinase inhibitors (TKIs) also cause such toxicity[1].The ability to predict drug-induced cardiotoxicity may reduce drug attrition and advance precision medicine.

Predictive modeling of adverse drug reactions (ADRs) by integrating information across databases and knowledgebase of biological activities, chemistry, and ADRs has been undertaken[2, 3, 4]. However, no predictive models of drug-induced cardiomyopathy utilizing signaling network information have been constructed. Harpaz et al[4] stressed the importance of harnessing multiple sources of knowledge, biological information and biomedical literature for predicting drug toxicity. In line with this notion, we reported herein predictive modeling by integrating prior knowledge, drug targets, and empirical data in order to enable the model to identify key predictors from a drug's mode of action, and to have the potential to inform lead identification and development.

To fill in the gap, we compiled a list of 31 toxic and non-toxic drugs that were transcriptomically profiled in human cardiomyocytes[5, 6, 7]; manually curated and compiled their clinical incidence of treatment-related cardiomyopathy; and conducted predictive modeling of drug-induced cardiomyopathy. Two predictive models were compared: 1) applying Elastic Net (EN) to gene expression data; 2) applying integer linear programming (ILP) to construct a drug's signaling network to reflect its mechanism of action[8], and then subjecting the nodes in individual drugs' signaling networks to EN regression . The ILP formulation[8] navigates a prior knowledge network of protein-protein, protein-TF, and TF-gene interactions, and identifies the pathways that connect a drug's targets to its DEGs. ILP not only optimizes the solution of finding a drug's signaling pathways but also enhances performance of predictive modeling by enabling identification of the subset of DEGs that are functionally relevant to a drug's mode of action. We further referenced literature for the microRNAs, which are reportedly of diagnostic value for heart failure and for drug-induced cardiomyopathy as

well as also regulate expression of our predictors, in hopes of shedding light on potential microRNAs as in-vivo drug-induced cardiomyopathy biomarkers.

Methods

Compilation of drugs and their clinical incidence of drug-induced cardiomyopathy

To compile the list of approved drugs that cause treatment-related cardiomyopathy, we referenced NIH's Common Terminology Criteria for Adverse Events (CTCAE 4.03)[9] and the Medical Dictionary for Regulatory Activities[10] for cardiomyopathy-related terms to text-mine approved drug labels. The terms used included cardiomyopathy, heart failure, congestive heart failure, cardiac failure, left ventricular dysfunction, left ventricular failure, and reduction in left ventricular ejection fraction. The current drug label PDF files (Drugs@FDA[1]) were processed using a text-mining analysis pipeline as published previously[11]. Individual rates of occurrence for cardiomyopathy were extracted by manual curation of drug labels, published redacted NDA reviews (Drugs@FDA), as well as published clinical studies.

Predictive modeling

Workflow and highlights of EN and ILP

As shown in Figure 1, we compared two pipelines of predictive modeling. For Pipeline 1, we applied EN to differentially expressed genes (DEGs) of a drug. For pipeline 2, we applied integer linear programming (ILP) to construct each drug's signaling pathway, and then subjected the genes/proteins in each drug's signaling network to EN regression.

EN is useful for predictive modeling when predictors greatly outnumber observations while simultaneously being able to identify statistically significant predictors[12]. EN regularization is useful for analyzing genomics of drug sensitivity in cancer[13].

We applied ILP to a drug's DEGs and protein targets to model its mode of action. These two levels of information are connected via signal transduction where the signal originates at drug targets, propagates intracellularly via a complex network of signaling cascades, passes through the layer of transcription factors (TF), and finally reaches the transcriptomic level of DEGs. We modeled the interactions in the knowledge network by using the logic formalism[14], which identified the minimum subset of the network to achieve the desired connectivity. We constructed the specific signaling network for each drug using an ILP formulation as published previously[8].

ILP will enhance predictive performance since it has the ability to capture cellular responses to a drug, to identify the subset of important functional DEGs, and to help differentiate between compounds and translate into improved performance.

Drug name normalization

Drug names were first normalized and identified by the PubChem compound identifier to ensure consistency when downloading data from Connectivity Map[15], DtoxS[5], STITCH[16, 17], and literature.

Compilation of drug targets

We compiled the targets of individual drugs from STITCH[16, 17], and the 'chemical-protein links' database and selected only human proteins. The proteins were identified by the SwissProt/EnsEMBL-identifier, and translated into HGNC gene symbols[18] using the R biomaRt package, in order to match with the nodes in the prior-knowledge network[19]. We used STITCH's 'interaction types for links' data file, from where we identified the drugs as activating or inhibiting individual target proteins. We used only those associating links between protein-drug pairs with an evidence score of >= 0.7.

Gene expression data sources and handling

Wherever data were available in Affymetrix probe IDs, the probe IDs (Affymetrix GeneChip Human Genome U133A Array) were translated into HUGO Gene Nomenclature Committee (HGNC) gene symbols[18] using the biomaRt package [20] and hgu133a2[21] packages in R, an open source statistical computing graphics systems.

Across all the gene lists we kept only those genes with fold change > 2 and p-value < 0.05 by a two-tailed, two-sample unequal variance Student's t-test, adjusted separately for the up and down gene lists with Bonferroni correction (p-value adjusted for multiple comparisons) [22].

A list of 75 drugs with drug-induced DEGs available from cancer cells[15] in Connectivity Map (CMap) were used for exploratory modeling (See Table S1 in Supplementary document-CMap). To conduct robust predictive modeling, we exhausted literature and databases and found a list of 31 drugs of which drug-induced DEGs in human cardiomyocytes[5] and stem cells-derived cardiomyocytes[5, 6, 7] were available. The 2 data sources for drug-induced perturbation of gene expression in cardiomyocytes were 1) 30 drugs from Drug Toxicity Signature Generation Center (DtoxS)[5], where PromoCells (primary human adult cardiomyocytes) were used, and 2) literature data of doxorubicin studied in human stem cells-derived cardiomyocytes[6, 7]. The size of each dataset was mainly constrained by the availability of DEGs data. For DToxS data, we downloaded the Level 2 gene expression data, calculated the fold changes, kept only those DEGs with a p-value < 0.05 and a fold change > 2, and merged them from different donors by averaging the fold changes while excluding any DEGs with opposite directions of fold change among donors.

Doxorubicin is widely studied for its dose-dependent cardiac toxicity, and is commonly dosed at 40mg-60 mg/m^2[21]. Following intravenous 60mg/m^2, its Cmax was 630ng/ml (1159nM) [23]. See Table S2 for a few studies of transcriptomic profiles of doxorubicin. For our modeling, we included the data from human-induced pluripotent stem cells-derived cardiomyocytes (hiPSC-CMs) by Chaudhari *et al*[7] and Burridge *et al*[6]. We included the gene expression data by Burridge et al [6] were at 100, 1000 and 10000nM and those by Chaudhari et al[7] were at 156nM (see Table S3 for the rationale).

Identifying a drug's mode of action using LP

We first built a prior-knowledge network as a scaffold for constructing a drug's signaling network by downloading from Reactome[19] the latest version (Version 2015) of the "Functional interactions derived from Reactome". As published previously[8], we

merged those interactions with transcription factors and obtained a network across the protein, transcription factor and gene levels, which contained 64,801 reactions, 2,585 signaling proteins and 12,376 genes. We applied ILP to optimize a drug's signaling network by providing as the input the scaffold mentioned above and its targets.

ILP formulation was solved using IBM ILOG CPLEX optimization studio[8] for the objective of optimizing a drug's network. Based on the constraints that mimic signal transduction[24] and adjustment to the specific case of very large (>10000 nodes) networks[8], the algorithm minimized the mismatch between the data of gene expression measurements and the prior knowledge pathway topology. The output was the optimal signaling network of a drug, identifying the molecular interactions that appeared to be functional based on the input of DEGs and drug targets. We were able to select the minimum part of a prior-knowledge network for each drug that could explain the data in hand. See Supplementary document-ILP for understanding the example of methotrexate signaling network captured by ILP (Fig S1) and how the proposed ILP formulation works.

Comparing predictive modeling by applying EN to a drug's DEGs vs to its ILP signaling network

To construct a matrix for EN regression, a drug was marked with 0 if classified as non-toxic and marked 1 if classified as toxic. We classified drugs by referencing approved labels for the criteria of "frequent adverse events being those occurring on one or more occasions in at least 1/100 patients; infrequent adverse events being those occurring in 1/100 to 1/1000 patients; rare events being those occurring in less than 1/1000 patients." Referencing the definition of rare events used in drug labeling and considering the distribution of clinical incidence, the number of drugs with gene expression data available, and the heterogeneity of clinical studies, we classified drugs into 2 classes, toxic for those with incidence $\geq$ 0.1% and non-toxic for those with incidence < 0.1%.

A column of "cardiotoxicity" was created with the clinical incidence score: 1 for "toxic" and 0 for "non-toxic". Each column corresponded to a single gene expressed in

at least one of the DEGs signatures. Individual DEGs of a drug were assigned a value of 1, -1 or 0 to reflect up-regulated, down-regulated or not reported, respectively (pipeline 1). The same assignments were applied to the nodes in each drug's ILP signaling network (pipeline 2).

In our modeling we used EN regression[12], and more specifically a linear regression model with an elastic net penalty determined using the R package glmnet[25]. EN regularization is defined by two parameters, alpha and lambda. EN regression is a mixing of LASSO and ridge regression and combines their two penalty terms for the alpha parameter. When alpha equals 0, EN performs as ridge regression and when alpha equals 1, EN performs as LASSO. In EN, the lambda parameter reflects shrinkage of the model's coefficients. When lambda equals 0, no shrinkage of the model's coefficients is performed but the coefficients decrease toward zero (though not exactly equal zero) as its value increases. We tried a range of values for alpha from 0 to 1 by a 0.01 step and selected the one that minimized the mean squared error. For that alpha value, we selected the value of lambda that gave the minimum mean cross-validated error.

To validate each model, we used Leave-One-Out Cross Validation (LOOCV) by leaving a drug's signature out one at a time (either DEGs or signaling network constructed from ILP) and did so across the whole list of drugs. Each time we calculated the accuracy, sensitivity and specificity for a predictive model, and selected and reported the model with the highest accuracy along with its precision, sensitivity and specificity. From the chosen predictive model, we extracted the predictors (genes/proteins) that best predicted drug-induced cardiotoxicity. The receiver operating characteristic (ROC) and precision-recall curves using the R package with the former plotted in smooth curve.

Pipeline 1-applying EN to DEGs

The results of 75 drugs with DEGs from CMap are summarized in Supplementary document-CMap. Among these 75 drugs with their DEGs from CMap, 24 drugs were toxic and the remaining 51 drugs were non-toxic.

A model matrix was constructed using cardiomyocyte data, with the 34 observations (toxicity classification) as rows and 15016 variables (gene expression) as columns. The predictive linear model was constructed by having as input all these variables for EN regularization. We tried all possible different cut-off scenarios (see the spreadsheet 'summary' of Table S4 for the results of the psccm_34_gen_heart trials and the detailed results of 18 models with different cut-offs in the spreadsheet '9'). For example, a cut-off of 10 meant that we ran the model by using only those genes that were expressed in at least 10 of the 34 signatures, meaning that the analysis started with 3508 genes while a cut-off of 15 started the analysis with the genes that appeared in at least 15 of the 34 signatures, meaning 464 genes were used as the cut-off.

Pipeline 2 – applying EN to gene/protein nodes in ILP-constructed signaling networks

We first performed exploratory modeling using a list of 75 drugs with gene expression data available in CMap and concluded that signaling networks of drugs derived from ILP outperformed their DEGs when applying EN regularization (see Figure S2 for ROC and precision-recall curves in Supplementary document-CMap).

We were able to find the ILP solutions for drugs with gene expression data in cardiomyocytes (Table S5) except cefuroxime, domperidone, and olmesartan. These three drugs were removed this modeling exercise. At the end, we had 31 signaling networks from 28 drugs (15 non-toxic drugs and 13 toxic drugs). See Table S5 for the gene/protein nodes in the signaling network of each individual drugs. We built a model matrix for the 31 signaling pathways/networks by using gene expression profiles from cardiomyocytes and by assigning 1 if a pathway node was up-regulated, -1 if it was down-regulated and 0 if it was not present in a drug's optimized signaling network. See the spreadsheet 'summary' of Table S4 for the results of the psccm_34_ILP_heart trial and the detailed results of 31 models with different cut-offs in the spreadsheet '10'.

Biological context of predictors

To gain translational insight, we searched literature for microRNAs that have been shown to be diagnostic markers of heart failure and also involved in regulation of gene expression. We mined literature and MiRTarBase, a database of experimentally-

validated microRNA-target interactions [26], for a list of microRNAs, which have been individually reported to regulate expression of our top gene/protein predictors, and also been reportedly detected in the circulation of heart failure patients with a varying degree of severity[27, 28] or of patients with doxorubicin-induced cardiomyopathy[29].

Results

The list of drugs and toxicity profile

The list of 31 drugs with their clinical profiles of treatment-related cardiomyopathy is summarized in Table 1. Literature search was also conducted to supplement clinical incidence of cardiomyopathy, if approved drug labels and published application reviews[1] did not have such information. Among the 31 drugs, there were 13 toxic drugs (41.9%) and there were 18 non-toxic drugs (59.1%). For those drugs without mention of cardiomyopathy-related toxicity described in their labels throughout the sections of clinical studies, post-marketing experiences, and warnings and precautions, we also searched literature and published reviews[1] to reach the conclusion that they are non-toxic drugs.

Predictive modeling

Applying EN to DEGs (pipeline 1)

Using Leave-One-Out Cross Validation (LOOCV) across the whole list of 30 drugs and their gene expression signatures, we achieved 79% accuracy and 75% precision, with 80% sensitivity and 79% specificity when using those genes that were expressed in at least 11 of the 34 signatures (a cut-off of 11 in spreadsheet '9' of Table S4). The results of elastic net regularization are shown in Figure 2 (2A, 2C), and the genes/proteins with non-zero coefficients are PHF19, HSPA8, RIF1, CD46, MXRA7, RAB27A, TOMM20, MYO6 and CCNA2. ROC curves and precision-recall curves are shown in Figure 3 and Figure S2 of Supplementary document-CMap), respectively.

Applying EN to the gene/protein nodes in ILP-constructed signalilng networks (pipeline 2)

By applying EN regression and LOOCV, we were able to increase both prediction accuracy and precision to 88%, with 88% sensitivity and 89% specificity, compared with the results from elastic net regression of DEGs (Table S4). EN regularization is shown in Figure 2 (2B and 2D). The result for the psccm_34_ILP_heart trial is in the spreadsheet 'summary' and the detailed results of 31 models with a cut-off ranging from 1 (5012 genes/proteins in at least 1 drug) to 31 (5 genes/proteins in at least 31 network signatures) are in the spreadsheet '10' of Table S4). The highest accuracy, sensitivity and specificity were achieved at cut-off of 10 with 189 genes/proteins from at least 10 drugs' signaling networks. ROC and precision-recall curves are shown in Figure 3 and Figure S3, respectively.

We concluded that EN-ILP (pipeline 2) outperformed EN alone (pipeline 1) when applied to the same set of DEGs.

Cardiac context of top predictors

Using EN regularization we were able to extract the protein/gene predictors that best predict the toxicity classification of drug-induced cardiotoxicity (either toxic for ≥ 0.1% clinical incidence or non-toxic for < 0.1%). The 33 protein/gene predictors along with their individual coefficients are summarized in Table 2. The network of the top 15 genes/proteins selected by the model is presented on Figure 4. Cardiac relevance of these predictors was reviewed and summarized in Table S6. The protein and gene predictors identified by EN-ILP reflected the key cellular biological factors for drug-induced cardiotoxicity. The EN regularization in our predictive modeling selected the protein/gene predictors that best predicted drug-induced cardiotoxicity.

We mined an evidence-based database of microRNAs[26] for those that reportedly regulate our top predictors, and also referenced literature to narrow the list to those that are reportedly of diagnostic value for heart failure. Summarized in Table 3 are our top 10 predictors and their individual regulating microRNAs that have reportedly been of diagnostic value for natural heart failure[27, 28] or for doxorubicin-induced cardiomyopathy[29].

Discussion

With the clinical incidence of drug-induced cardiomyopathy as a dependent variable, ILP-enhanced predictive modeling increased prediction accuracy from 79% to 88%, compared to modeling with EN and DEGs alone. This improved prediction signified the ability of ILP to computationally capture a drug's mode of action through constructing its signaling pathways for the purpose of predictive modeling. ILP offers the advantage of integrating our prior knowledge of biological protein interactions and drug targets (Reactome, STITCH), transcription factors, and DEGs into predictive modeling. ILP also optimizes the size of a drug's network signaure in addition to capturing the signaling pathways of a drug. Take lapatinib as an example, it had 2265 DEGs from cardiomyocytes while from this set of DEGs, its ILP network consisted of 1923 nodes including its targets, proteins involved in its signaling transduction, transcription factors, and functional DEGs.

The 33 gene/protein predictors along with their individual positive or negative coefficients could be used to predict 'toxic' or 'non-toxic' for a drug by linear summation using their individual levels of expression (either up-regulation (+) or down-regulation (-)) from its ILP-constructed signaling network. The predictive power of this systems pharmacology predictive model will increase with the amount of data in the training set.

Among the 31 drugs used to conduct predictive modeling, the distribution of toxic (13) vs non-toxic (18) classification was acceptable, though not ideal. Among them, there were 18 kinase inhibitors (17 tyrosine kinase inhibitors (TKIs) and 1 serine/threonine kinase inhibitor), which might seemingly be off-balance from the perspective of the diversity of drug class. Vemurafenib is a serine/threonine kinase inhibitor and not toxic. The distribution of toxic (8) and non-toxic (9) drugs among the 17 TKIs was acceptable. TKIs in general lack target specificity, have multiple targets, and were designed to disrupt the signaling pathways that are vital to cancer cell survival[30]. Unfortunately, several of these signaling pathways also play a critical role in cardiomyocyte biology[31]; consequently, several TKIs impair cardiac function. Within this context, our predictive modeling could be useful for predicting cardiac toxicity for future new chemical entities.

All top 15 gene/protein predictors have relevant cardiac functions except ZNF 823 (Supplementary Table 5). Interesting, CYP3A4 was an important predictor. Though CYP3A4 does not have biological interactions with other predictors, as shown in Figure 4, it is a major drug metabolizing enzyme[1]. Among the 31 drugs, 10 of 13 (85%) toxic drugs and 11 of 18 (61%) non-toxic drugs were metabolized by CYP3A4. The toxic drugs, that are primarily or extensively metabolized by CYP3A4, included amiodarone, axitinib, cytrabine, dasatinib, doxorubicin, imatinib, ponatinib, sorafenib, sunitinib and vandetanib[1, 32, 33, 34]. For non-toxic drugs, they are bosutinib, crizotinib, cyclosporine, domperidone, erlotinib, gefitinib, lapatinib, regorafenib, ruxolitinib, tofacitinib, and ursodeoxycholic acid[1, 35].

Some top predictors are biologically associated with focal adhesion kinase (FAK), a non-receptor protein-tyrosine kinase, which is involved in cytoskeleton-associated network of signaling proteins[36]. Focal adhesion complexes play a critical role in how cultured cardiomyocytes respond to mechanical and neurohormonal stimuli, and in development of heart failure[37]. FAK activation plays a role in the adaptive response to cardiac afterload and in myocyte growth via the AKT/mTOR pathway[38]. FAK cleavage is mediated by CASP3 family during apoptosis of human normal cells[39], and occurs with activation of EPHA2 and p38 MAPK during doxazosin-induced apoptosis of a cardiac cell line[40]. FAK activates STAT1 during cell attachment[41], and plays a role in cell migration with one of its actions being associated with PDGFR signaling complex[42]. In short, the top predictors are important to maintain normal cardiac function.

Per literature, some microRNAs that reportedly regulated expression of our predictors have also been shown to be of diagnostic value for heart failure with a varying degree of severity (Table 3)[27, 28]. Among them, miR193-3p and miR26b-5p reportedly regulated more predictors than other microRNAs, and regulated 4 and 3 of our top predictors, respectively. It might be worthy of clinical studies to determine whether miR193-3p and miR26b-5p are useful in-vivo biomarkers for drug-induced cardiomyopathy. Literature search uncovered a recent study that investigated circulating microRNAs in children with anthracycline-induced acute heart injury[29]. Elevated miR-

29b and miR-499 in the circulation seemed to correlate with troponin elevation in these children, and were identified as potential cardiomyopathy biomarkers[29]. This observation of miR-29b elevation in doxorubicin-induced cardiomyopathy differed from an observation of decreased expression of miR-29b-3p in the coronary sinus blood of heart failure patients[28]. MiR-29b-3p regulates expression of one of our top 10 predictors, PDGFRA. Further studies are needed to investigate the role of miR-29b in drug-induced cardiomyopathy or in natural heart failure. Even though miR-27b reportedly regulated CYP3A4[26, 43], literature search did not uncover any reports that suggested miR-27b be of diagnostic value for drug-induced cardiomyopathy.

Integrating clinical incidence with the modes of action of a drug, which is depicted as its signaling network, for predictive modeling is a strength of our study. There are, however, some limitations in our approach: 1) non-toxic slightly outnumbered toxic drugs, 2) limitation of ILP where no biological feedback controls are considered and assumptions adopted in ILP formulation, 3) DEGs of doxorubicin in cardiomyocytes were from different sources than the rest of 30 drugs, and 4) availability of transcriptomic profiling data in cardiomyocytes. Furthermore, our study inherited the shortcomings associated with the databases and knowledge base used for our modeling. The impact of disease indications on the incidence and severity of treatment-related cardiomyopathy is not well characterized.

Our predictive modeling of integrating clinical incidence of drug-induced cardiomyopathy with the signaling network of toxic and non-toxic drugs not only is useful for further improving its predictive power, but also identifies important gene/protein predictors that have relevant cardiac biological functions. Above all, the top genes/protein predictors are reportedly regulated by specific microRNAs that have been shown to be of diagnostic value for heart failure or drug-indued cardiomyopathy. These predictors might be useful for shedding light on potential microRNAs as in-vivo biomarkers of drug-induced cardiomyopathy.

Study Highlights

What is the current knowledge on this topic?

There is no translational predictive modeling that integrates a drug's mode of action with clinical observation of toxicity.

What question did this study answer?

This study addresses the question of 1) how to conduct systems pharmacology predictive modeling that integrates the modes of action of drugs and their clinically observed occurrence of treatment related cardiomyopathy.

What does this study add to our knowledge?

This study adds to the knowledge of 1) the proteins/genes that are top predictors of drug-induced cardiomyopathy, 2) utility of drugs' modes of action in the form of signaling pathways for predicting drug-induced cardiomyopathy.

How might this change drug discovery, development, and/or therapeutics?

This study enables pharmaceutical scientists to further translational systems pharmacology modeling to facilitate development of therapeutics.

Author Contributions: D.E.M. and J.P.F.B. wrote the manuscript; D.E.M. and J.P.F.B. designed the research; D.E.M., J.G.H., and N.V. performed the research; I.N.M. and L.G.A. contributed new reagents/analytical tools.

References

1. Drugs@FDA. *http://wwwaccessdatafdagov/scripts/cder/drugsatfda/indexcfm*. (Accessed July 2016)
2. Pouliot Y, Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin Pharmacol Ther* **90** 90-99. (2011)
3. Liu M*, et al.* Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc* **19** e28-35. (2012)
4. Harpaz R*, et al.* Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther* **91** 1010-1021. (2012)
5. DtoxS - Drug Toxicity Signature Generation Center - Data & Resources. *<https://martip03uhpcmssmedu/dataphp>*. (Accessed Feb, 2017)
6. Burridge PW*, et al.* Human induced pluripotent stem cell-derived cardiomyocytes recapitulate the predilection of breast cancer patients to doxorubicin-induced cardiotoxicity. *Nat Med* **22** 547-556. (2016)
7. Chaudhari U*, et al.* Identification of genomic biomarkers for anthracycline-induced cardiotoxicity in human iPSC-derived cardiomyocytes: an in vitro repeated exposure toxicity approach for safety assessment. *Arch Toxicol* **90** 2763-2777. (2016)
8. Melas IN*, et al.* Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury. *Integr Biol (Camb)* **7** 904-920. (2015)
9. Institute NC. CTCAE 4.03 (Common Terminology Criteria for Adverse Events) *https://ctepcancergov/protocoldevelopment/electronic_applications/ctchtm* (Accessed July 2016)
10. ICH. MeDRA (Medical Dictionary for Regulatory Activities). *http://wwwmeddracom/*. (Accessed July 2016)
11. Hur J, Guo AY, Loh WY, Feldman EL, Bai JP. Integrated systems pharmacology analysis of clinical drug-induced peripheral neuropathy. *CPT Pharmacometrics Syst Pharmacol* **3** e114. (2014)
12. Zou HaH, T. . Regularization and variable selection via the elastic net. *Journal of Royal Statitical Society Series B* **B 67** 301-320. (2005)
13. Institute WTS. Genomics of Drug Sensitivity in Cancer. *http://wwwcancerrxgeneorg/news*. (accessed August, 2017)
14. Melas IN, Samaga R, Alexopoulos LG, Klamt S. Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Comput Biol* **9** e1003204. (2013)
15. Connectivity Map (CMAP). *https://wwwbroadinstituteorg/connectivity-map-cmap*. (accessed August, 2016)
16. Szklarczyk D*, et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* **44** D380-384. (2016)
17. CONSORTIUM S. Stitch-EMBL. *http://stitchemblde/*. (accessed August, 2016)
18. (EMBL-EBI) EBI. HUGO Gene Nomenclature Committee. *http://wwwgenenamesorg/*. (Accessed August, 2016)
19. Reactome. *http://wwwreactomeorg/*. (Accessed August, 2016)
20. Gentleman R. annotate: Annotation for microarrays. . *R package version 1500*. (Accessed August, 2016)
21. Carlson M. hgu133a2.db: Affymetrix Human Genome U133A 2.0 Array annotation data (chip hgu133a2). . *R package version 323*. (Accessed August, 2016)
22. Computing RFfS. R: A language and environment for statistical computing. *https://wwwr-projectorg/*. (2016)

23. Barpe DR, Rosa DD, Froehlich PE. Pharmacokinetic evaluation of doxorubicin plasma levels in normal and overweight patients with breast cancer and simulation of dose adjustment by different indexes of body mass. *Eur J Pharm Sci* **41** 458-463. (2010)
24. Mitsos A*, et al.* Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Comput Biol* **5** e1000591. (2009)
25. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33** 1-22. (2010)
26. Chou CH*, et al.* miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* **44** D239-247. (2016)
27. Schulte C, Westermann D, Blankenberg S, Zeller T. Diagnostic and prognostic value of circulating microRNAs in heart failure with preserved and reduced ejection fraction. *World J Cardiol* **7** 843-860. (2015)
28. Marques FZ, Vizi D, Khammy O, Mariani JA, Kaye DM. The transcardiac gradient of cardio-microRNAs in the failing heart. *Eur J Heart Fail* **18** 1000-1008. (2016)
29. Leger KJ*, et al.* Circulating microRNAs: Potential Markers of Cardiotoxicity in Children and Young Adults Treated With Anthracycline Chemotherapy. *J Am Heart Assoc* **6**. (2017)
30. Karaman MW*, et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol* **26** 127-132. (2008)
31. Force T, Krause DS, Van Etten RA. Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. *Nat Rev Cancer* **7** 332-344. (2007)
32. Fabre G, Julian B, Saint-Aubert B, Joyeux H, Berger Y. Evidence for CYP3A-mediated N-deethylation of amiodarone in human liver microsomal fractions. *Drug Metab Dispos* **21** 978-985. (1993)
33. Zientek MA*, et al.* In Vitro Kinetic Characterization of Axitinib Metabolism. *Drug Metab Dispos* **44** 102-114. (2016)
34. Scripture CD, Figg WD. Drug interactions in cancer therapy. *Nat Rev Cancer* **6** 546-558. (2006)
35. Bodin K, Lindbom U, Diczfalusy U. Novel pathways of bile acid metabolism involving CYP3A4. *Biochim Biophys Acta* **1687** 84-93. (2005)
36. Schlaepfer DD, Hauck CR, Sieg DJ. Signaling through focal adhesion kinase. *Prog Biophys Mol Biol* **71** 435-478. (1999)
37. Samarel AM. Focal adhesion signaling in heart failure. *Pflugers Arch* **466** 1101-1111. (2014)
38. Clemente CF*, et al.* Focal adhesion kinase governs cardiac concentric hypertrophic growth by activating the AKT and mTOR pathways. *J Mol Cell Cardiol* **52** 493-501. (2012)
39. Grossmann J*, et al.* Hierarchical cleavage of focal adhesion kinase by caspases alters signal transduction during apoptosis of intestinal epithelial cells. *Gastroenterology* **120** 79-88. (2001)
40. Jehle J*, et al.* Regulation of apoptosis in HL-1 cardiomyocytes by phosphorylation of the receptor tyrosine kinase EphA2 and protection by lithocholic acid. *Br J Pharmacol* **167** 1563-1572. (2012)
41. Xie B*, et al.* Focal adhesion kinase activates Stat1 in integrin-mediated cell migration and adhesion. *J Biol Chem* **276** 19512-19523. (2001)
42. Sieg DJ*, et al.* FAK integrates growth-factor and integrin signals to promote cell migration. *Nat Cell Biol* **2** 249-256. (2000)
43. Pan YZ, Gao W, Yu AM. MicroRNAs regulate CYP3A4 expression via direct and indirect targeting. *Drug Metab Dispos* **37** 2112-2117. (2009)

Figure Legends

Figure 1 Workflow of predictive modeling. We built datasets using gene expression data and we compared two piplelines to predict clinical drug-induced cardiomyopathy and extract features that best predict such toxicity. Running the Gene Expression Data at hand through a linear regression model with elastic net regularization or constructing signaling networks from the data before modeling using an ILP formulation.

Figure 2 Plots of elastic net regularization results. Panels A and B show selection of the alpha parameter in the elastic net regularization by minimizing the leave one out cross validation mean squared error to extract the features (genes) that best predict clinical incidence of cardiomyopathy. Panels C and D show the number of variables kept in the model, with a vertical line showing the optimal number for maximization of accuracy. Panels A and C refer to the results of analyzing gene expression data only while panels B and D correspond to the results of analyzing drugs' signaling networks obtained from integer linear programming formulation analysis. Each of the plotted lines in Figure 2C and 2D corresponds to a variable (for example a specific gene's expression) and shows how its coefficient changes with the log lambda parameter of elastic net. The vertical line shows the optimal number of parameters kept and their coefficients for maximization of accuracy.

Figure 3 Receiver operating characteristic (ROC) curves. Panel A: ROC curve from modeling DEGs using Elastic Net (EN) and panel B: ROC curve from modeling by 1) subjecting these DEGs to integer linear programming to construct their individual drugs' signaling networks and then subject these networks to EN.

Figure 4 Interactions among Top 15 gene/protein predictors. Interactions among the top 15 genes/proteins selected by our model to best predict cardiomyopathy using cardiomyocytes data are depicted as a network using Stitch website for visualization. Small nodes correspond to protein of unknown 3D structure and large nodes to known or predicted. Edges represent protein-protein associations and the intensity of the line is proportional to the confidence score of each association. The confidence score is calculated by combining the probabilities from all evidence channels and is corrected for random observation probability.

Table Titles

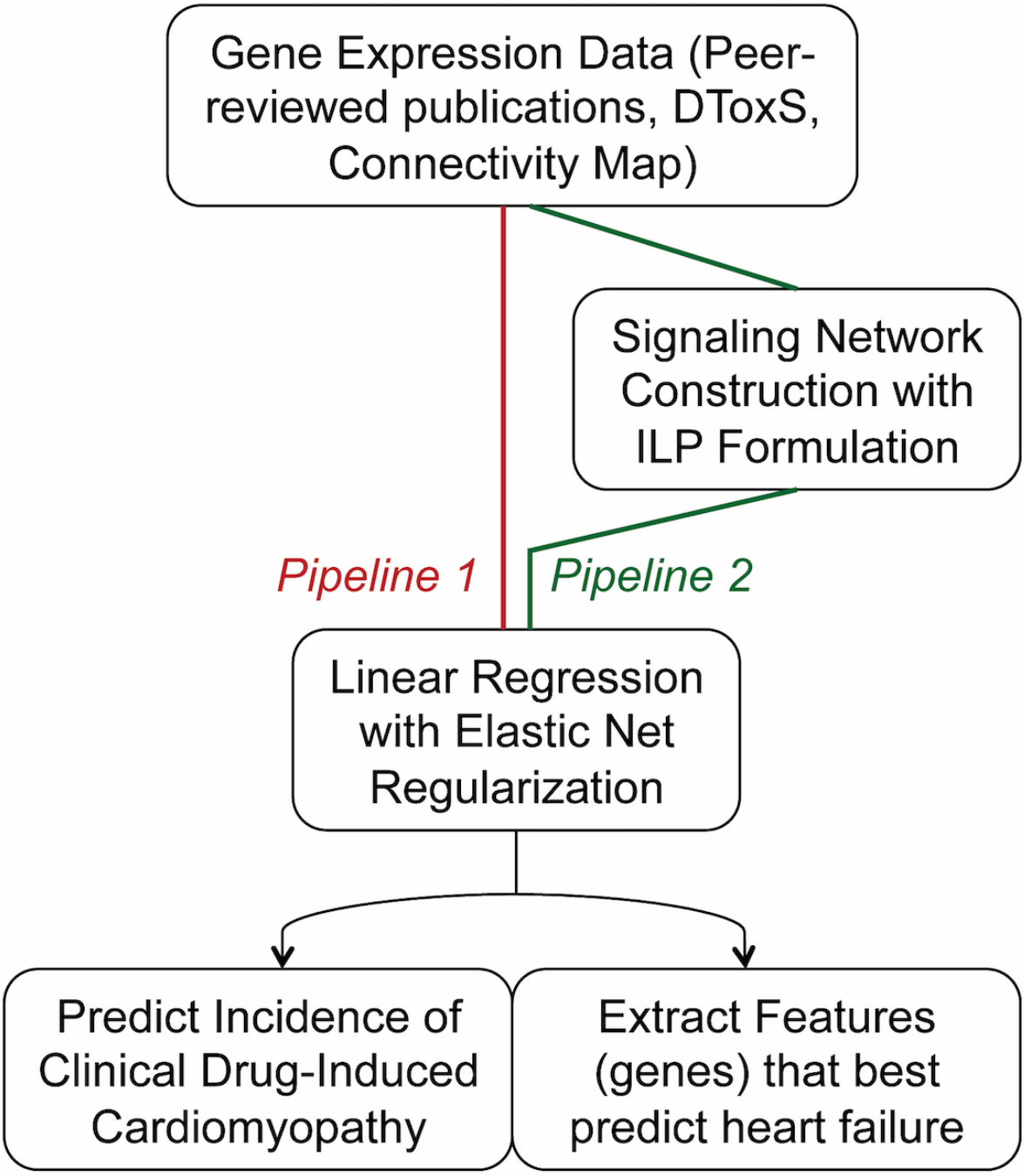
Table 1. The list of drugs with gene expression in cardiomyocytes and their cardiomyopathy classification

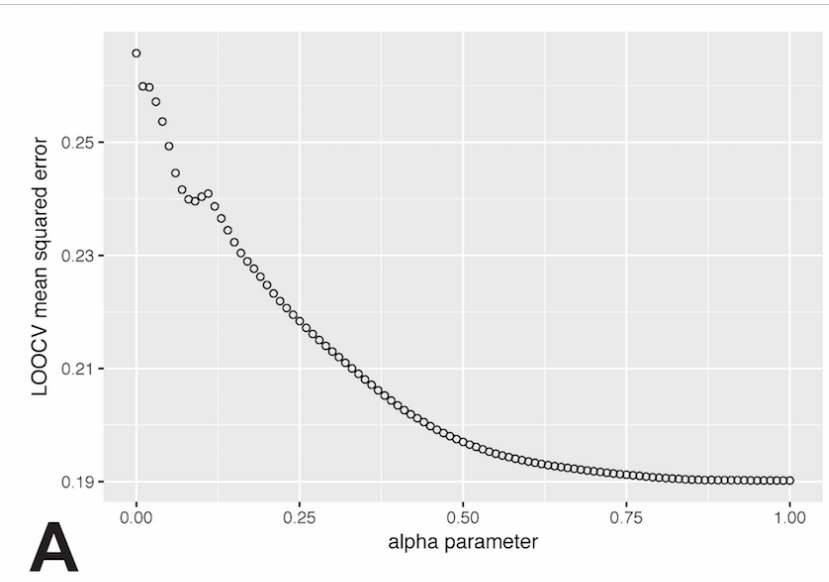
Table 2. Predictors with non-zero coefficients from modeling/analysis of cardiomyocyte data

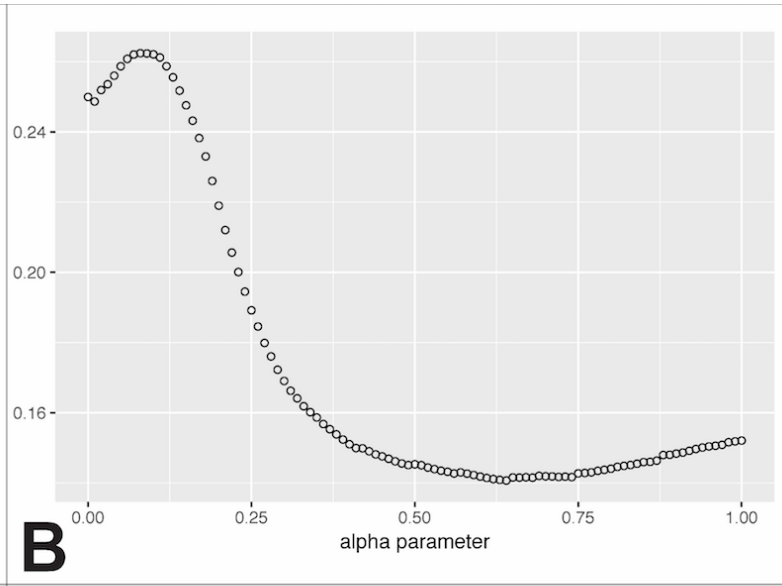
Table 3. Top 10 predictors and their corresponding regulating microRNAs that are reportedly of diagnostic value for heart failure

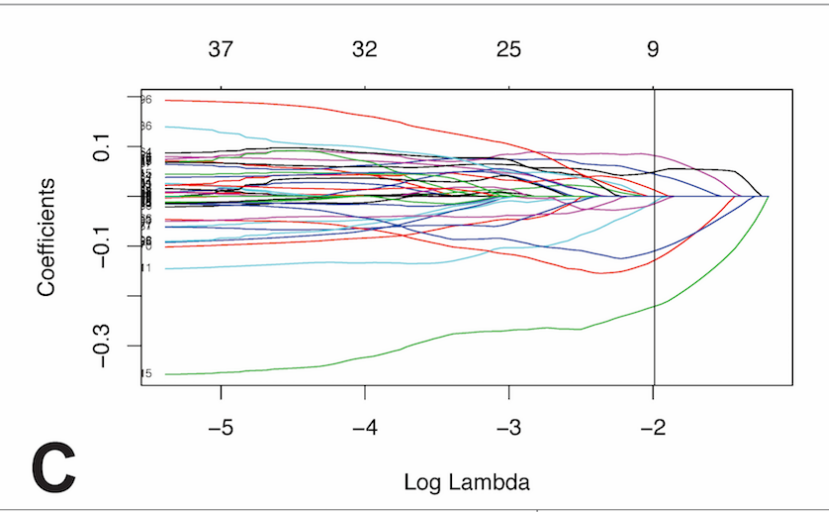
File Titles

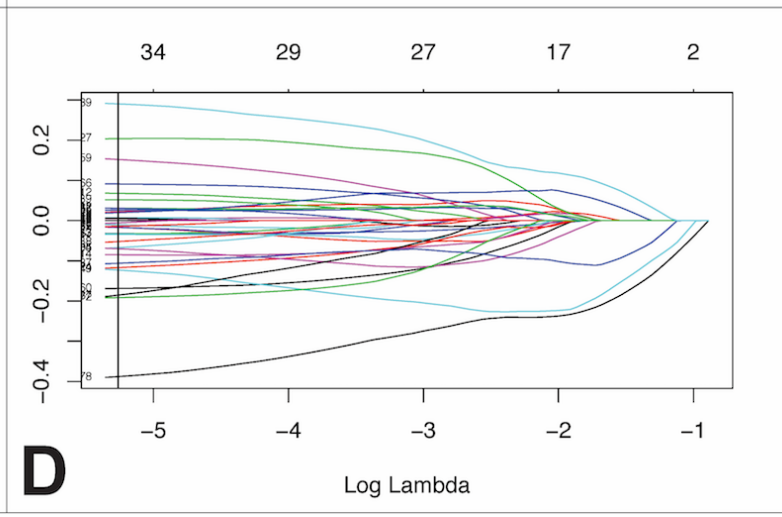
1. Supplementary document _CMap: Supplementary documentation of Cmap data: Predictive modeling using cancer cell lines data from Connectivity Map. This document contains Table S1 (the list of 75 drugs) and Figure S2.
2. Supplementary document-ILP: a drug's mode of action illustrated by ILP. This document contains Figure S1 (the methotrexate signaling network).
3. Supplementary Figure S3. Panels A and B are the precision/recall curves for the 34 differentially expressed genes modeled using Elastic Net (EN) and an integer linear programming-EN combination, respectively.
4. Table S2: a few studies of transcriptomic profiles of doxorubicin
5. Table S3: Rationale for including 4 DEGs profiles of doxorubicin
6. Table S4: Detailed results of predictive modeling
7. Tabel S5: ILP solutions of 34 DEGs profiles.
8. Table S6. Biological functions of the top 15 gene/protein predictors

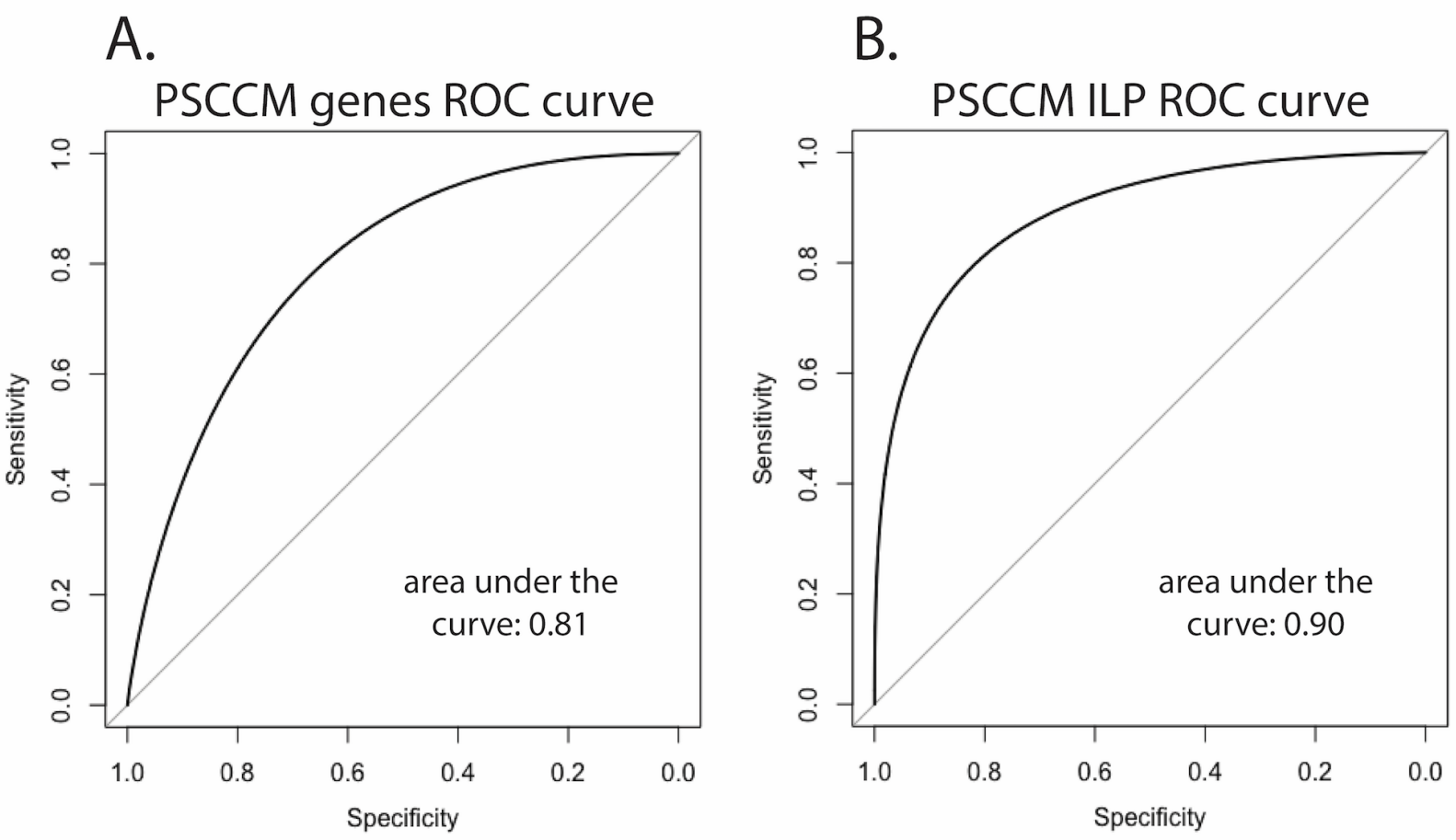












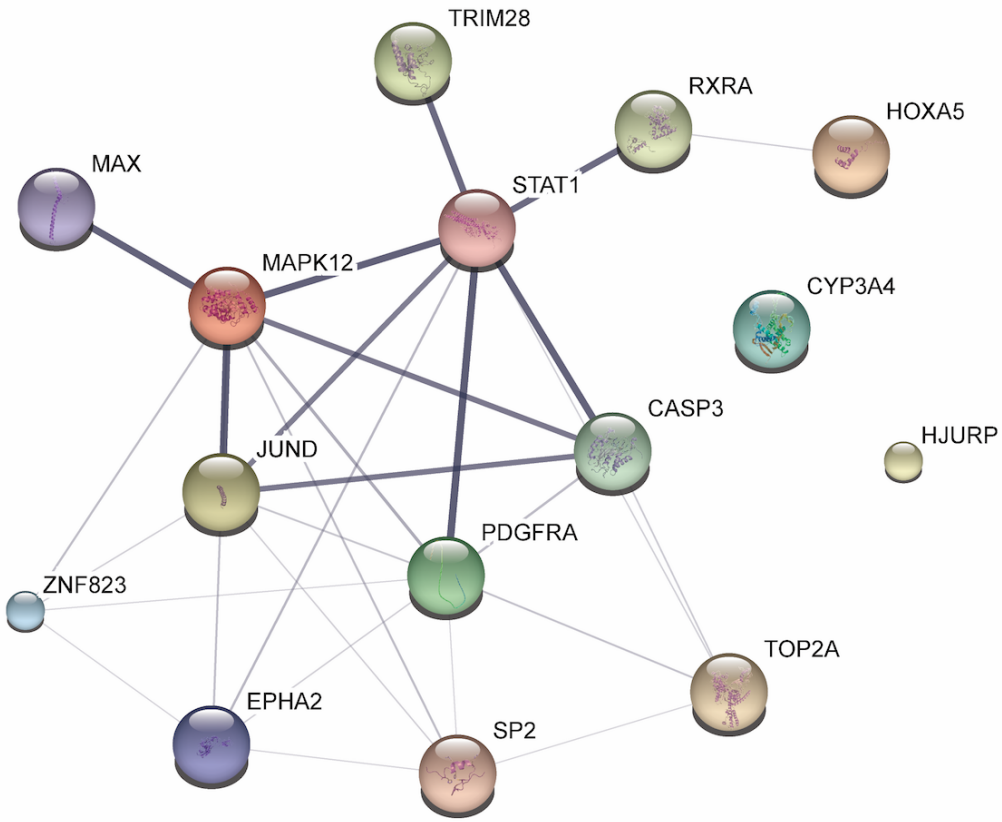


Table 1. The list of drugs with gene expression in cardiomyocytes and their cardiotoxicity classification

drug name	Classification	Reference*
Afatinib	0	Drugs@FDA & literature search
Alendronate	0	Drugs@FDA & literature search
Amiodarone	1	Drugs@fda
Axitinib	1	Drugs@fda
Bosutinib	0	Drugs@FDA & literature search
Cefuroxime	0	Drugs@FDA & literature search
Crizotinib	0	Drugs@FDA & literature search
Cyclosporine	0	Drugs@FDA & literature search
Cytarabine	1	NIH DailyMed
Dasatinib	1	Drugs@FDA
Diclofenac	1	Drugs@FDA ; Coxib and traditional, Nsaid Trialists' Collaboration[52]
Domperidone	0*	Not approved by US FDA
Doxorubicin	1	Drugs@FDA
Diethylpropion	0	Drugs@FDA & literature search
Erlotinib	0	“
Gefitinib	0	“
Imatinib	1	Drugs@FDA
Lapatinib	0	Drugs@FDA; Perez et al.[53]
Methotrexate	0	Drugs@FDA & literature search
Olmesartan	0	Drugs@FDA & literature search
Paroxetine	1	Drugs@FDA
Ponatinib	1	Drugs@FDA
Regorafenib	0	Drugs@FDA & literature search
Ruxolitinib	0	Drugs@FDA & literature search
Sorafenib	1	Drugs@FDA
Sunitinib	1	“
Tofacitinib	0	Drugs@FDA & literature search
Trametinib	1	“
Ursodeoxycholic Acid	0	Drugs@FDA & literature search
Vandetanib	1	Drugs@FDA
Vemurafenib	0	Drugs@FDA & literature search

Note: toxic: 1 (clinical incidence ≥ 0.1%), non-toxic: 0 (clinical incidence <0.1%).
https://dailymed.nlm.nih.gov/dailymed/
*: Domperidone was profiled by DtoxS and toxicity information was from http://www.hc-sc.gc.ca/dhp-mps/medeff/reviews-examens/domperidone-eng.php

Table 2. Predictors with non-zero coefficients from modeling/analysis of cardiomyocyte data

Gene/protein	coefficient	Gene/protein	coefficient	Gene/protein	coefficient	Gene/protein	coefficient
CYP3A4	-0.39	TOP2A	-0.11	FLI1	-0.03	H2AFX	-0.01
ZNF823	0.29	MAX	0.09	TCF12	-0.03	IRF1	-0.011
CASP3	0.20	JUND	-0.08	AHR	0.03	MAP3K5	0.01
HJURP	-0.19	MAPK12	-0.07	BCR	0.03	E2F1	0.01
EPHA2	-0.19	RXRA	0.07	GATA3	0.03	SMOC2	0.01
STAT1	-0.17	HOXA5	-0.07	SMC3	0.02	CYP2D6	-0.01
SP2	0.15	STAT5A	-0.05	EDN1	0.02		
PDGFRA	-0.12	TCF7L2	0.05	FOXF2	-0.02		
TRIM28	-0.12	NR4A2	-0.03	CTCFL	-0.02		

*: Nodes from drugs' signaling networks constructed using integer linear programming (ILP) included proteins (targets and protein-protein interactions) and genes (differentially expressed). The gene/protein nodes from ILP were then subjected to elastic net regularization.

Table 3. Top 10 predictors and their corresponding regulating microRNAs that are reportedly of diagnostic value for heart failure

Predictors	Regulating microRNAs* that are of diagnostic value	References
CYP3A4	No information	
ZNF823	miR193-3p (↓)	Schulte et al [28]
CASP3	miR-375**, miR-26b-5p(↓); miR-30e-5p(↓), let-7a-5p (↑)	Schulte et al [28]; Marques et al[29]
HJURP	miR-671-5p(↑)	Schulte et al [28]
EPHA2	miR-26b-5p(↓), miR-193b-3p(↓); miR-16-5p(↓)	Schulte et al [28]; Marques et al [29]
STAT1	miR 145-5p (↓)	Schulte et al [28]
SP2	miR-29a-3p (↓), miR-638**	Schulte et al [28]
PDGFRA	miR-140-5p (↓); miR-26b-5p (↓);miR-29b-3p(↓);181a-5p (↑); miR-1233 (↑)	Schulte et al [28]; Marques et al [29]
TRIM28	miR-423-5p (inconsistent reports), miR-193b-3p (↓), miR-183-3p (↓),miR-92a-3p (↓)	Schulte et al [28]
TOP2A	miR-193b-3p (↓), miR-21-5p (↑)	Schulte et al [28], Marques et al [29]

Note: Regulating microRNAs are from Chou et al[27] (http://mirtarbase.mbc.nctu.edu.tw); **: differentiating heart failure with reduced ejection fraction from heart failure with preserved ejection fraction. ↑and ↓ represent elevation and decrease, respectively, compared to healthy controls.